Canon
Delighting You Always

Understanding the Query: THCIB and THUIS at NTCIR-10 Intent Task

Yunqing Xia[1] and Sen Na[2]
[1] Tsinghua University
[2] Canon Information Technology (Beijing) Co. Ltd.

Before we start

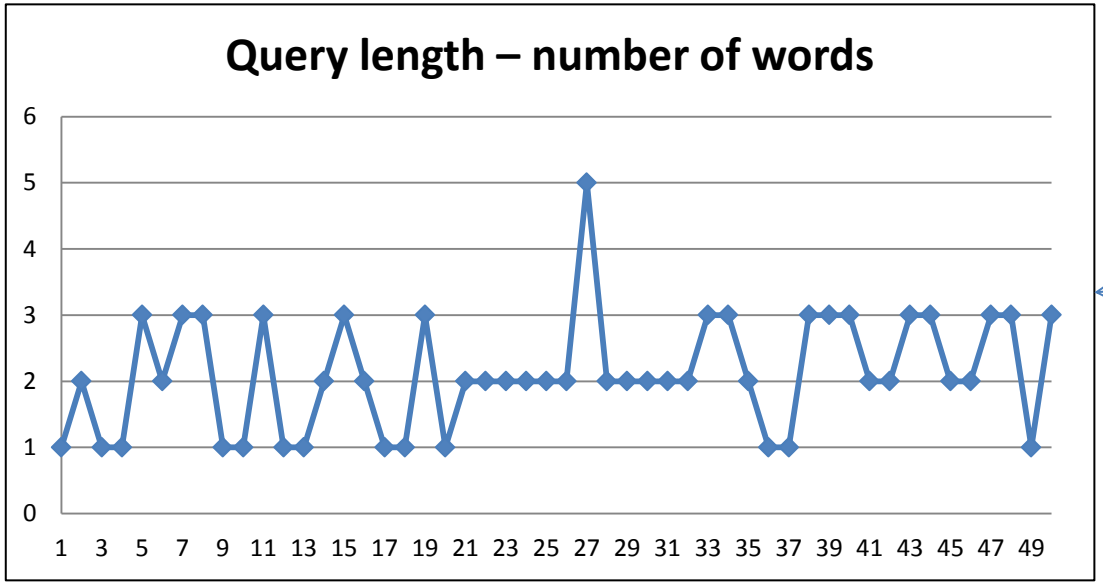
- Who are we?
  - **THUIS** is the research team at Intelligent Search group at Center for Speech and Language Technology, Tsinghua University
  - **THCIB** is the joint research team between THUIS and Canon Information Technology (Beijing) Co. Ltd..
- Why did we participate NTCIR INTENT task?
  - We believe intent mining is one of the most promising technologies to make the search engines smarter thus more helpful to human.
  - We view query-based intent mining as a major topic in our research group
- What is task/subtask we participated?
  - Subtopic mining: Systems are required to return a ranked list of ***subtopic strings*** in response to a given topic query while the top N subtopic strings should be ***both relevant and diversified*** as much as possible.

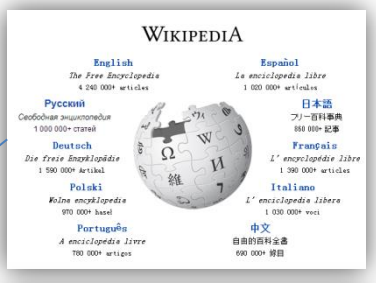
Outline

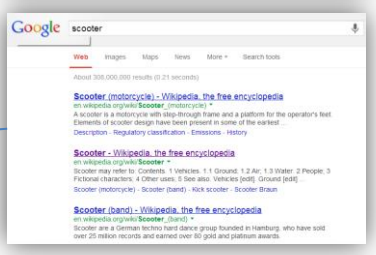
- The motivation
- System overview
- What make our system different?
- Evaluation
  - The submitted runs
  - Results and discussion
- Conclusion and future work

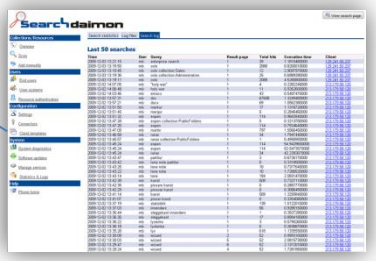
The Motivation (1/3)

- ISSUE #1: **Query is usually very short**









- SOLUTION #1: Appling BIGGER CONTEXT in query understanding
  - General knowledge base: Wikipedia
  - User behavior data: Query log, search engine auto-completions and suggestions
  - Search results: Title and snippet

The Motivation (2/3)

- ISSUE #2: **Subtopic surface strings are redundant**

furniture for small spaces store
{furniture for small spaces market
{furniture for small spaces wholesale
{furniture for small spaces shop
{furniture for small spaces center
……

furniture for small spaces store {store, market, wholesale, shop, center, …}

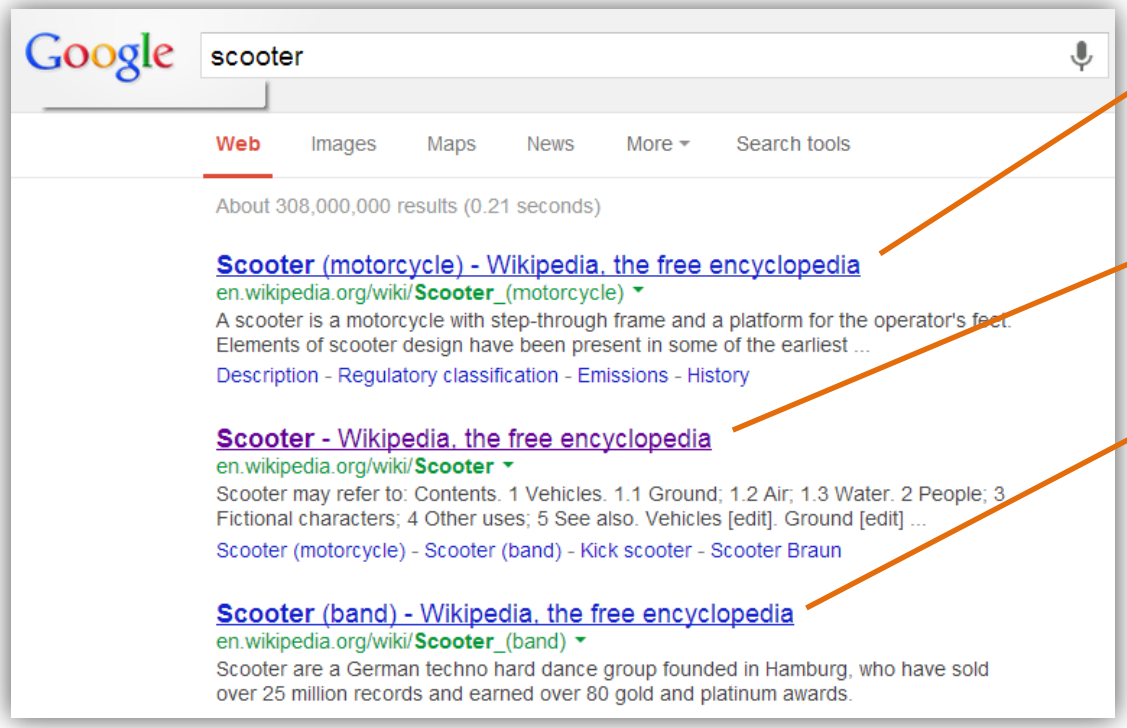
{furniture for small spaces Tokyo
furniture for small spaces New York
{furniture for small spaces London
{furniture for small spaces Hong Kong
{furniture for small spaces Indonesia
……

furniture for small spaces Tokyo {Tokyo, New York, London, Hong Kong, Indonesia, …}

- SOLUTION #2: Discover the implicit intents by clustering the subtopic surface strings
  - A sense-based clustering algorithm

The Motivation (2/3)

- ISSUE #3: **Relevance is no longer effective** for intent ranking

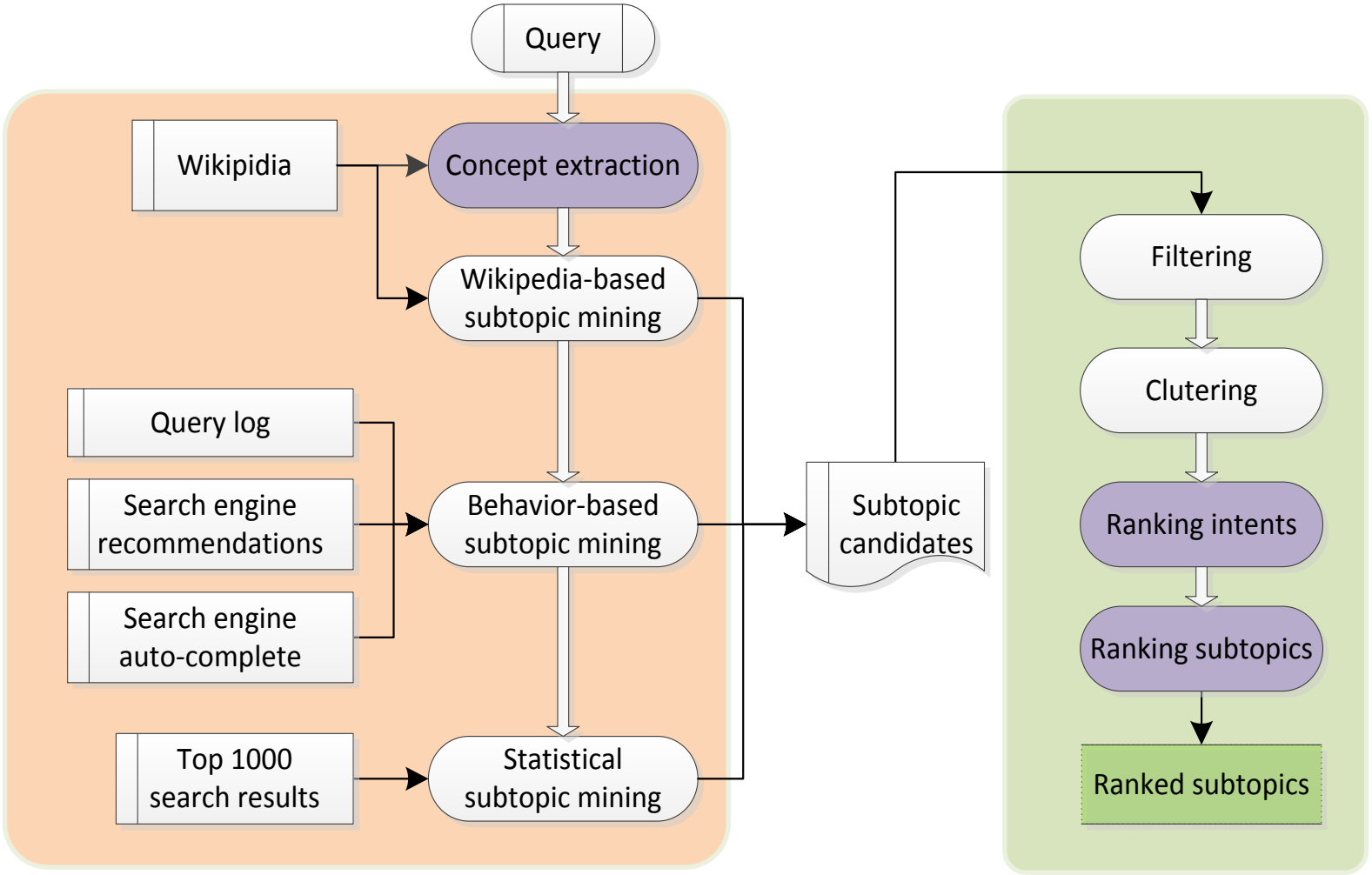


- SOLUTION #3: Ranking intents considering both relevance and diversity
  - A unified intent weighting model and a subtopic selecting strategy

Outline

- The motivation
- System overview
- What make our system different?
- Evaluation
  - The submitted runs
  - Results and discussion
- Conclusion

System overview



Subtopic candidate mining (SCM)

Subtopic candidate ranking (SCR)

Outline

- About the NTCIR10 INTENT-2 task
  - Who are we?
  - Why do we participate NTCIR INTENT task?
  - Task/subtask we participated
- The motivation
- System overview
- What make our system different?
- Evaluation
  - The submitted runs
  - Results and discussion
- Conclusion and future work

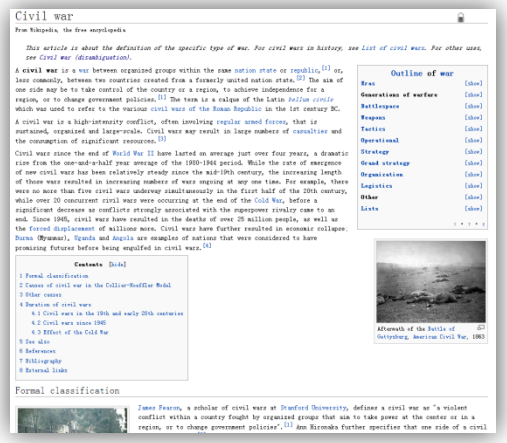
What make our system different?

- Concept based
  - Wikipedia entries and related entries
  - From query analysis to expansion
  - From subtopic extraction to intent mining
  - From relevance to diversity
  - From weighting to ranking
- Discovering intent for diversification
  - Word sense induction
  - Intent induction/disambiguation
  - Entity analysis to address homogeneous exclusive subtopics

SCM: Extracting concepts from query

- Downloading the entire Wikipedia
  - Entry ==> Concept
  - Concept ==> Definition
  - Concept ➔ Related concepts
- Bi-directional maximum entry matching
- Using the multiple matches in the disambiguation page
- Using redirects when no entry is exactly matched

"battles in the civil war" ➔ "battle", "civil war"

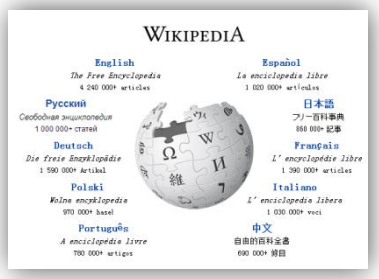


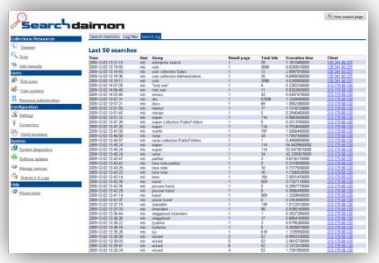
SCM: Query expansion

- Wikipedia
  - Synonymous entries (redirects) and the related concepts
  - Polysemous entries (disambiguation pages)
- Intent schema
  - {concepts, prepositions, wild cards}
  - "hobby store": "* of hobby store", "* at hobby store", "hobby store in *", "hobby store at *", etc.
- Concept repositioning
  - "battles in civil war" ➔ "battles civil war", "civil war battles"

- The motivation:
  - Reforming the query so as to obtain subtopic candidates as many as possible (in query auto-completion, query suggestions, etc.)

SCM: Extracting subtopic candidates





- Wikipedia – general knowledge base
  - Concept definition
- User Behavior Data – user centric data
  - Co-occurrence
  - Search engine tools (auto-completion, query suggestion)
- Search Results – pseudo feedback
  - Query topics (word senses) within snippets of top N=1000 results

Wikipedia concept definition

subtopic

- **Ground**
  - Kick scooter
  - Motorized scooter
  - Scooter (motorcycle)
  - Knee scooter
  - Mobility scooter
  - Eccentric-hub scooter
  - Square scooter
- **Air**
  - Douglas A-4 Skyhawk
  - Air Scooter
- **Water**
  - Underwater scooter
  - Water scooter
  - Ice boat
- People
  - Scooter Braun
  - Lloyd L. Burke (nicknamed "Scooter")
  - Dill Stokes (nicknamed "Scooter")
- Fictional characters
  - Scooter (comics)
  - Scooter (*Coronation Street*)
  - Scooter (Gobots)
  - Scooter (Muppet)
  - Scooter (*SpongeBob SquarePants*)
  - Scooter (talking baseball)
  - *Scooter: Secret Agent*

Scooter

From Wikipedia, the free encyclopedia
(Redirected from Scooters)

Scooter may refer to:

Vehicles [edit]

Ground [edit]

- Kick scooter, a vehicle propelled by a standing rider pushing off the ground
- Motorized scooter, a motorized version of a kick scooter
- Scooter (motorcycle), a motorcycle with a step-through frame and a platform for the feet
- Knee scooter, a mobility device used for walking by people with leg injuries
- Mobility scooter, a motorized chair
- Eccentric-hub scooter, a two-wheeled vehicle propelled by a bouncing rider
- Square scooter, a square plank with four swivel casters

Air [edit]

- Douglas A-4 Skyhawk, a ground-attack aircraft
- AirScooter, a theoretical ultralight helicopter

Water [edit]

- Underwater scooter, a piece of diving equipment
- Water scooter, a recreational watercraft
- Ice boat, a vehicle for quick travel across water, ice or snow

People [edit]

- Scooter Braun (born 1981), American talent manager
- Lloyd L. Burke (1924–1999), U.S. Army soldier, nicknamed "Scooter"
- Oll Trigg (born 1950), American political figure
- Dill Stokes (1917–2007), American baseball player, nicknamed "The Scooter"

Fictional characters [edit]

- Scooter (comics)
- Scooter (*Coronation Street*)
- Scooter (Gobots)
- Scooter (Muppet)

User Behavior Data

- The user search log (e.g., ClueWeb09)
- Tools of commercial search engines based on user behavior data
  - Auto-completion
  - Query suggestion
- With expanded queries based on concepts

Search results

- Search with concept as a whole keyword
  - In query <battles in the civil war>, <"civil war"> is one keyword *WORD*
  - In Web pages, <"civil war"> is one keyword *WORD* ('war' must immediately follow 'civil')
- Induce aspects of the query using WSI (word sense induction) technique
  - LDA + keyword extraction
  - Labeled LDA
  - *Sense based LDA: a sense based clustering algorithm*

Reference: A paper submitted to CIKM 2013.

SCR: Re-calculating the relevance score

- Replacing bag of word with bag of Wikipedia concepts
  - BM25 again.
- Incorporating source score

$$p_t = w_{ST}(t) + w_{SC}(t)$$

- w_st(t): Relevance score of the subtopic candidates
- w_sc(t): Importance score (empirical) of the source where the subtopic comes from.

SCR: Discovering intents

- Clustering subtopics candidates with Affinity Propagation (AP) algorithm
  - Calculating subtopic similarity with VSM-based cosine similarity
  - Extraction concept-based VSM features from snippets of the top 50 search results with subtopic string as a query.
  - Choosing mean of the similarity matrix as clustering preference value
- The revised version
  - Choosing mean of the subtopic importance value (=relevance + resource weight)

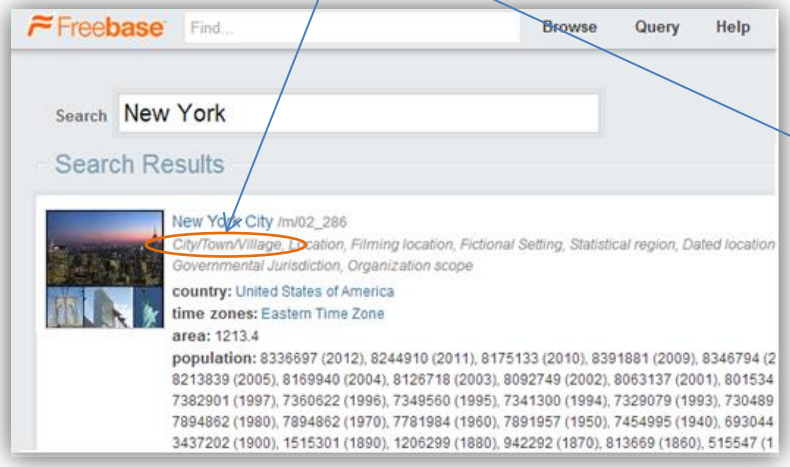
SCR: Weighting the intents

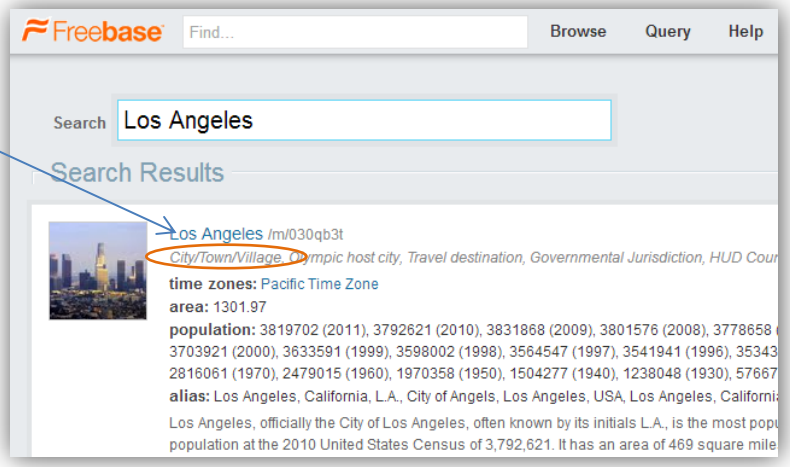
- A simple sum equation

$$w_{IN} = \sum_{i=1}^{N} [w_{ST}(t_i) + w_{SC}(t_i)]$$

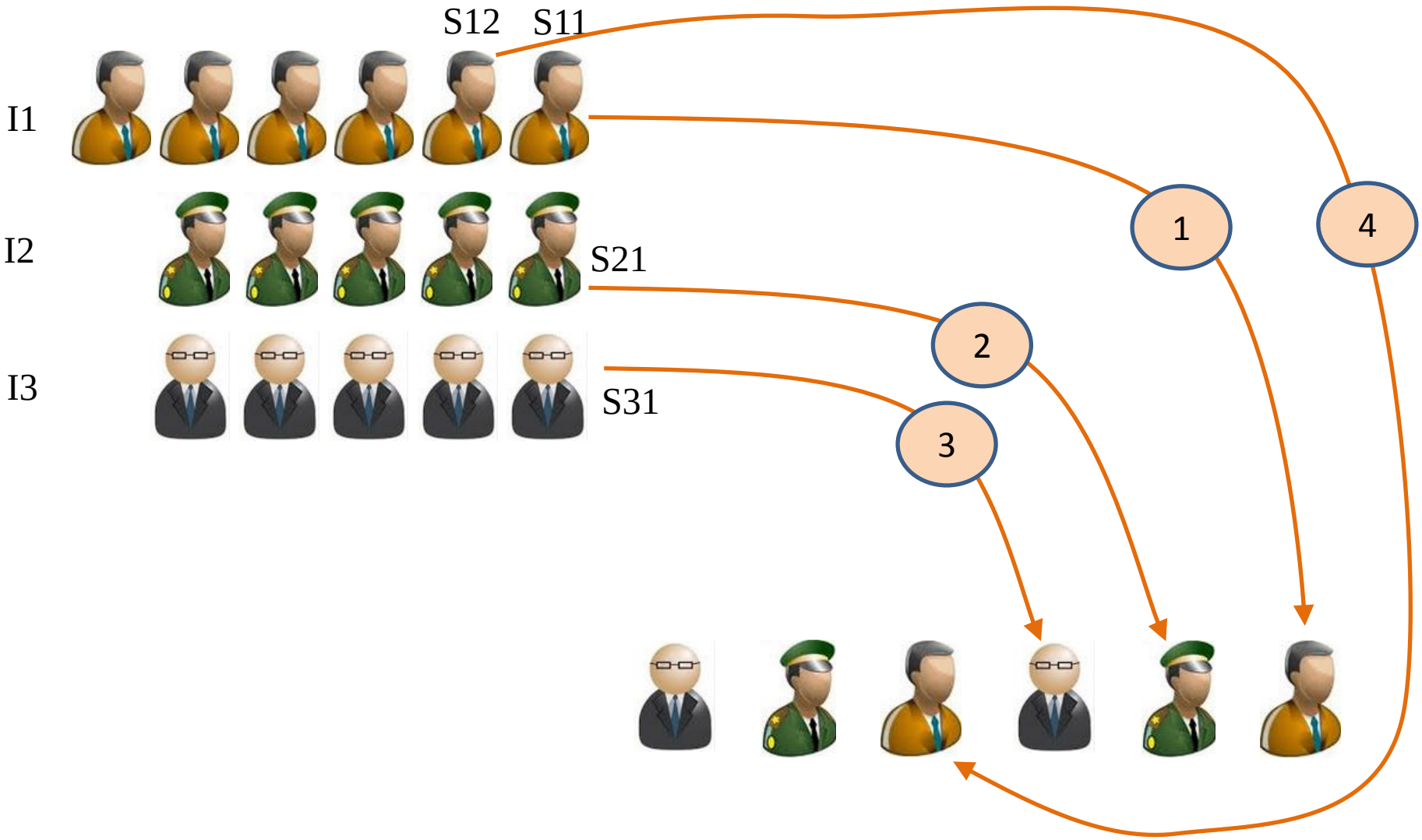
SCR: Entity analysis

- Homogenous exclusive entities are found many in subtopic candidates
  - "furniture for small spaces New York"
  - "furniture for small spaces Los Angeles"
- Freebase - a global resource of ontology
  - It provides HTTP API for data retrieval
  - The whole dump data can be downloaded from Web
- Judgment of homogenous exclusive entities
  - Sharing the same immediate father node!
  - "City/Town/Village"





SCR: Selecting for ranking



Outline

- The motivation
- System overview
- What make our system different?
- Evaluation
  - The submitted runs
  - Results and discussion
- Conclusion and future work

The submitted runs

- We submitted 5 runs for English task

RUN ID	Description
THCIB-S-E-1A	SCM (1.Concept extraction + 3.Wikipedia + 4.Query log+ 5.Search results + 6.Filtering) + SCR (similarity + source importance+ relevance)
THCIB-S-E-2A	THCIB-S-E-1A + SCM (2.Query expansion)
THCIB-S-E-3A	THCIB-S-E-2A +SCR (4.Entity analysis)
THCIB-S-E-4A	THCIB-S-E-3A +SCR (3.Intent discovering with standard AP + 5.Intent weighting+ 6. Subtopic selecting)
THCIB-S-E-5A	THCIB-S-E-4A + SCR (3.Intent mining with revised AP)

- We submitted 4 runs for Chinese task
  - No Freebase in Chinese (Run 3 in English is not planned for Chinese task).

Results and discussion – Performance

- Rank:
  - Run 2> Run 1> Run 3> Run 5> Run 4
- Observations
  - Concept-based query expansion is useful in subtopic mining (Run 2 vs. 1)
  - Entity analysis is not appropriately used (Run 4)
  - Performance of intent discovery can be improved (Run 3)
  - Intent-based subtopic weighting model can be improved (Run 3)

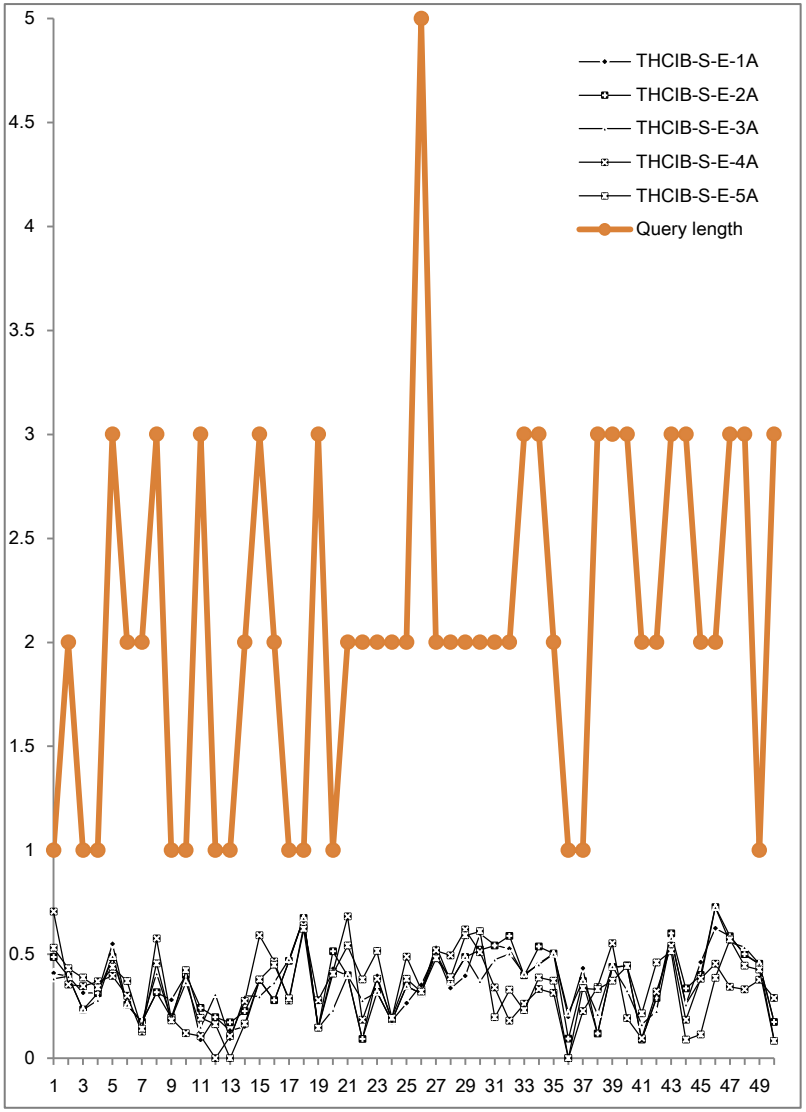
English Subtopic Mining runs

cut-off	run name	I-rec	D-nDCG	D#-nDCG
@10	THCIB-S-E-1A	0.3785	0.3384	0.3584
	THCIB-S-E-2A	**0.3797**	**0.3499**	**0.3648**
	THCIB-S-E-3A	0.3681	0.3383	0.3532
	THCIB-S-E-4A	0.3502	0.3323	0.3413
	THCIB-S-E-5A	0.3662	0.3215	0.3438
@20	THCIB-S-E-1A	0.5769	0.3274	0.4522
	THCIB-S-E-2A	**0.5899**	**0.3406**	**0.4653**
	THCIB-S-E-3A	0.5544	0.3251	0.4397
	THCIB-S-E-4A	0.477	0.2784	0.3777
	THCIB-S-E-5A	0.5395	0.304	0.4218
@30	THCIB-S-E-1A	**0.693**	0.3177	**0.5054**
	THCIB-S-E-2A	0.6743	**0.3284**	0.5014
	THCIB-S-E-3A	0.6486	0.3244	0.4865
	THCIB-S-E-4A	0.5855	0.2691	0.4273
	THCIB-S-E-5A	0.6339	0.2986	0.4662

- Performance in Chinese task is similar

Results and discussion – Per-topic analysis

- Best runs on the 50 queries
  - THCIB-S-E-1A 8
  - THCIB-S-E-2A 13
  - THCIB-S-E-3A 6
  - THCIB-S-E-4A 13
  - THCIB-S-E-5A 10
  - No run is consistently best, and each shows strength (further study is necessary)
- Query length
  - Our system is not sensitive to query length (Num. of words)
  - Other factors should be studied.



Outline

- The motivation
- System overview
- What make our system different?
- Evaluation
  - The submitted runs
  - Results and discussion
- Conclusion and future work

Conclusion and future work

- Conclusion
  - Incorporating concepts and word senses in subtopic mining and ranking brings marginal performance gain (NLP is positive to SM).
  - Subtopic ranking based on the automatically discovered intent is promising (**though** more work is required to improve intent quality).
- Future work
  - Deeper understanding the query: better subtopic extraction and intent discovery
  - Complexity issue: concept based indexing and retrieval
  - How about navigational and transactional query?

Acknowledgement

We thank Canon Inc. for supporting this research (No. TEMA2012).

We also thank the valuable comments from INTENT-2 organizer.

Canon
Delighting You Always

THANK YOU!
Q&A

We also welcome offline discussion by sending emails to

yqxia@tsinghua.edu.cn